Published online 7 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D815–D820***
doi:10.1093/nar/gkr916

YMDB: the Yeast Metabolome Database

Timothy Jewison[1], Craig Knox[1], Vanessa Neveu[1], Yannick Djoumbou[2], An Chi Guo[1], Jacqueline Lee[1], Philip Liu[1], Rupasri Mandal[1], Ram Krishnamurthy[1], Igor Sinelnikov[1], Michael Wilson[1] and David S. Wishart[1,2,3,*]

[1]Department of Computing Science, [2]Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G 2E8 and [3]National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, AB, T6G 2M9 Canada

Received August 15, 2011; Revised October 6, 2011; Accepted October 8, 2011

ABSTRACT

The Yeast Metabolome Database (YMDB, http://www.ymdb.ca) is a richly annotated 'metabolomic' database containing detailed information about the metabolome of *Saccharomyces cerevisiae*. Modeled closely after the Human Metabolome Database, the YMDB contains >2000 metabolites with links to 995 different genes/proteins, including enzymes and transporters. The information in YMDB has been gathered from hundreds of books, journal articles and electronic databases. In addition to its comprehensive literature-derived data, the YMDB also contains an extensive collection of experimental intracellular and extracellular metabolite concentration data compiled from detailed Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR) metabolomic analyses performed in our lab. This is further supplemented with thousands of NMR and MS spectra collected on pure, reference yeast metabolites. Each metabolite entry in the YMDB contains an average of 80 separate data fields including comprehensive compound description, names and synonyms, structural information, physico-chemical data, reference NMR and MS spectra, intracellular/extracellular concentrations, growth conditions and substrates, pathway information, enzyme data, gene/protein sequence data, as well as numerous hyperlinks to images, references and other public databases. Extensive searching, relational querying and data browsing tools are also provided that support text, chemical structure, spectral, molecular weight and gene/protein sequence queries. Because of *S. cervesiae's* importance as a model organism for biologists and as a biofactory for industry, we believe this kind of database could have considerable appeal not only to metabolomics researchers, but also to yeast biologists, systems biologists, the industrial fermentation industry, as well as the beer, wine and spirit industry.

INTRODUCTION

Metabolomics is a field of 'omics' research that is primarily focused on the identification and characterization of small molecule metabolites in cells, organs and organisms (1). Together with genomics, transcriptomics and proteomics these four 'omics' disciplines form the cornerstones to systems biology. However, relative to its more mature 'omics' cousins, metabolomics still lags far behind in developing or formalizing its software and database infrastructure (2). This is because the needs of metabolomics researchers span a very diverse range of scientific disciplines including organic chemistry, analytical chemistry, biochemistry, molecular biology and systems biology. In other words, metabolomics requires a tight blending of the tools found in both bioinformatics and cheminformatics. To address these informatics challenges, we (and others) have been steadily developing a set of comprehensive and open access tools to lay a more solid software/database foundation for metabolomics (2–4). In particular, our group has developed several widely used organism- or discipline-specific databases including the Human Metabolome Database (HMDB) (5), DrugBank (6), the CyberCell database (CCDB) (7), the Toxin/Toxin-Target database (T3DB) (8) and the Small Molecule Pathway Database (SMPDB) (9). HMDB, T3DB, DrugBank and SMPDB were specifically developed to address the metabolomics, toxicology, pharmacology and systems biology associated with humans (i.e. *Homo sapiens*), whereas CCDB was specifically developed to address the metabolomics and systems biology needs for *Escherichia coli*.

We believe that the establishment and maintenance of organism-specific metabolomics databases is absolutely

*To whom correspondence should be addressed. Tel: +780 492 0383; Fax: +780 492 1071; Email: david.wishart@ualberta.ca

critical to the field of metabolomics as each organism has a unique and chemically distinct metabolome. The 'naïve' identification of metabolites, by simple mass matching for instance, without regard to their origin (organism or man-made) frequently leads to spurious, humoros or meaningless compound identifications (10). Therefore, as part of our ongoing effort to create species-specific metabolomic resources for other model organisms we have now turned our attention to yeast, or more specifically, *Saccharomyces cerevisiae*.

The metabolic byproducts of *S. cerevisiae* fermentation are particularly interesting from both a biochemical and an industrial point of view. Indeed, *S. cerevisiae* (and its various strains) is perhaps the world's most important microbial biofactory, playing a key role in industrial chemical or biofuel production (ethanol), in the baking industry, as well as in beer, wine and spirit production. Together, these yeast-based industries are worth more than one trillion dollars per year to the global economy (11). As a model organism for molecular biologists, *S. cerevisiae* is certainly the most intensively studied microbe and perhaps the most well understood living thing on earth. Being one of the first organisms to be fully sequenced (12) and being particularly amenable to unique and powerful genetic manipulations (13,14) the sequence, function and interacting partner(s) of every gene/protein in *S. cerevisiae* is now almost completely known. This knowledge is contained in a number of excellent yeast-specific resources including SGD (15), YPD (16), CYGD (17) and FunSpec (18). This remarkably detailed molecular knowledge has also made *S. cerevisiae* a favorite model organism for systems biologists, leading to the development of some very useful resources aimed at modeling or describing yeast pathways and metabolism including YeastNet (19), MetaCyc (20), KEGG (21) and Reactome (22). Each of these excellent databases contains valuable information on primary yeast metabolic reactions, pathways and primary yeast metabolites.

Unfortunately, none of these systems biology databases contains information on the secondary metabolites of yeast fermentation (those compounds that give wine, beer and certain cheeses or breads their flavor or aroma), yeast-specific lipids, yeast volatiles or yeast-specific ions. These actually represent hundreds of industrially and biochemically important compounds. Furthermore, none of today's current set of yeast systems biology databases provides detailed metabolite descriptions, intra- or extracellular concentrations, growth conditions, physicochemical properties, subcellular locations, reference Nuclear Magnetic Resonance (NMR) or Mass Spectrometry (MS) spectra or other parameters that might typically be needed by researchers interested in yeast metabolism or yeast fermentation. For metabolomics researchers, as well as industrial chemists working with yeast byproducts, these kinds of data need to be readily available, experimentally validated, fully referenced, easily searched and readily interpreted. Furthermore, they need to cover as much of the yeast metabolome as possible. In an effort to address these shortcomings with existing yeast systems biology databases and to create a database specifically targeting the needs of yeast metabolomics, we have developed the Yeast Metabolome Database (YMDB).

DATABASE DESCRIPTION

The YMDB is a combined bioinformatics–cheminformatics database with a strong focus on quantitative, analytic or molecular-scale information about yeast metabolites and their associated properties, pathways, functions, sources, enzymes or transporters. The YMDB builds upon the rich data sets already assembled by such resources as YeastNet 4.0 (19), MetaCyc (20), KEGG (21), UniProt (23), ChEBI (24) and HMDB (5). But it also brings in a large body of independently collected literature data, as well as a significant quantity of experimental data, including NMR spectra, MS spectra and validated metabolite concentrations, to compliment this electronic or literature-derived data.

The diversity of data types, the quantity of experimental data and the required breadth of domain knowledge made the assembly of the YMDB both difficult and time-consuming. To compile, confirm and validate this comprehensive collection of data, more than a dozen textbooks, several hundred journal articles, nearly 30 different electronic databases and at least 20 in-house or web-based programs were individually searched, accessed, compared, written or run over the course of the past 18 months. The team of YMDB contributors and annotators included analytical chemists, NMR spectroscopists, mass spectroscopists and bioinformaticians with dual training in computing science and molecular biology/chemistry.

The YMDB currently contains more than 2000 yeast metabolite entries that are linked to nearly 27 000 different synonyms. These metabolites are further connected to some 66 non-redundant pathways and 916 reactions involving 857 distinct enzymes and 138 transporters. More than 750 compounds are also linked to experimentally acquired 'reference' ^{1}H and ^{13}C NMR and MS/MS spectra. Concentration data (intracellular and extracellular) is also provided for a total of 627 compounds. The complete collection of data in the YMDB occupies a total of 1.1 GB. Relative to other yeast metabolite/pathway databases, YMDB is substantially larger and significantly more comprehensive. A detailed comparison of YMDB to other widely known yeast resources is provided in Table 1.

The YMDB is modeled closely after the HMDB. As a result, it has many of the features found in the HMDB including efficient, user-friendly tools for viewing, sorting and extracting metabolites, proteins, pathways or chemical taxonomy information (Figure 1). These are available through the YMDB navigation bar (located at the top of every YMDB web page) that lists seven pull-down menu tabs ('Home', 'Browse', 'Search', 'About', 'Help', 'Download' and 'Contact Us'). To further aid in navigation and searching, nearly every viewable page in the YMDB, including the 'Home' page, supports simple text queries through a text search box located near the top of each YMDB web page. This text

Table 1. Comparison of the size and content of different yeast-specific or yeast-containing metabolism/metabolomics databases

Database Content	YMDB	YeastNet 4[a]	MetaCyc	KEGG
Number of metabolites	2007	792	688	720
Number of data fields	81	14	19	12
Number of NMR spectra	1540	0	0	0
Number of MS spectra	1346	0	0	0
Number of external database hyperlinks	15	4	3	5
Concentrations	Yes	No	No	No
Compound descriptions	Yes	No	No	No
Cell locations	Yes	Yes	No	No
Pathways	Yes	No	Yes	Yes
Sequence search	Yes	No	Yes	Yes
Structure search	Yes	No	No	Yes
Molecular weight search	Yes	No	Yes	No
NMR spectral search	Yes	No	No	No
MS spectral search	Yes	No	No	No
Chemical taxonomy	Yes	No	Yes	No

[a]YeastNet 4.0 reports a total of 1494 metabolites but only 792 are unique.

search tool, which can be specified to search through either protein or metabolite data fields, supports text matching, accommodates mis-spellings and highlights the text where the word is found. A more advanced text search that supports Boolean constructs and permits more precise data field specifications is also available.

In addition to these extensive text search capabilities, the YMDB also offers general database browsing via the 'Browse' buttons located in the YMDB menu bar. Five different Browsing options are available including Metabolite Browse (for viewing and sorting metabolites), Protein Browse (for viewing and sorting proteins), Reaction Browse (for viewing chemical reactions), Pathway Browse (for viewing yeast-specific KEGG pathways) and Class Browse (for viewing groups of compounds by their chemical taxonomy or class). Each of the Browsing views is presented as a set of navigable/sortable synoptic summary tables. These tables are, in turn, linked to more detailed 'MetaboCards' and 'ProteinCards' similar to those found in DrugBank and HMDB. Clicking on a MetaboCard or ProteinCard button opens a web page describing the compound or protein of interest in much greater detail. Every MetaboCard entry contains >50 data fields devoted to chemical or physico-chemical data and synoptic biological data (names, sequences, accession codes). Each ProteinCard entry contains >30 data fields devoted to biochemical, nomenclature, gene ontology and sequence data for metabolically important yeast enzymes and transporters. In addition to providing comprehensive numeric, sequence and textual data, each MetaboCard and ProteinCard also contains hyperlinks to many other databases (KEGG, BioCyc, PubChem, ChEBI, PubMed, PDB, UniProt, GenBank), abstracts, references, digital images and applets for viewing molecular structures.

Adjacent to the 'Browse' menu, the 'Search' menu offers nine different querying tools including Chem Query, Text Query, Sequence Search, Data Extractor, MS Search, MS/MS Search, GC/MS search, NMR Search and 2D NMR Search. Chem Query is YMDB's chemical structure search utility. It can be used to sketch (through ChemAxon's freely available chemical sketching applet) or paste a Simiplified Molecular Input Line Entry Specification (SMILES) string (25) of a query compound into the Chem Query window. Submitting the query launches a structure similarity search that looks for common substructures from the query compound that matches the YMDB's database of known yeast compounds. Users can also select the type of search (exact or Tanimoto score) to be performed. High scoring hits are presented in a tabular format with hyperlinks to the corresponding MetaboCards. The Chem Query tool allows users to quickly determine whether their compound of interest is a known yeast metabolite or chemically related to a known yeast metabolite. In addition to these structure-similarity searches, the Chem Query utility also supports compound searches on the basis of molecular weight ranges.

YMDB's sequence searching utility (Sequence Search), which supports both single and multiple sequence queries allows users to search through YMDB's collection of 1104 known enzymes, transporters and other target proteins. With Sequence Search, gene or protein sequences may be searched against YMDB's sequence database by pasting the FASTA formatted sequence (or sequences) into the Sequence Search query box and pressing the 'submit' button. A significant hit reveals, through the associated MetaboCard hyperlink, the name(s) or chemical structure(s) of metabolites that may act on that query protein. With Sequence Search metabolite–protein interactions from newly sequenced yeast species or strains may be readily mapped via the *S. cerevisiae* data in the YMDB.

YMDB's data extraction utility (Data Extractor) employs a simple relational database system that allows users to select one or more data fields and to search for ranges, occurrences or partial occurrences of words, strings or numbers. The data extractor uses clickable web forms so that users may intuitively construct SQL-like queries. Using a few mouse clicks, it is relatively simple to construct complex queries ('find all metabolites that are substrates of alcohol dehydrogenase and have boiling points above 80°C') or to build a series of highly customized tables. The output from these queries can be provided in HTML format with hyperlinks to all associated MetaboCards or as an easily downloaded comma separate value file.

YMDB's NMR and MS search utilities allow users to upload peak lists and to search for matching compounds from the database's collection of MS and NMR spectra. The YMDB currently contains 1540 experimentally obtained ^{1}H and ^{13}C NMR spectra (with spectral collection conditions) for 466 different compounds (most collected in water at pH 7.0, 10 mM for ^{1}H, 50 mM for ^{13}C) measured in our lab or obtained from the BioMagResBank (BMRB) (26). Most of the NMR spectra are fully assigned. It also contains 951 MS/MS (Triple-Quad) spectra for 317 pure compounds analyzed by our laboratory. An additional 400 MS or MS/MS spectra were obtained from MassBank (27). The YMDB

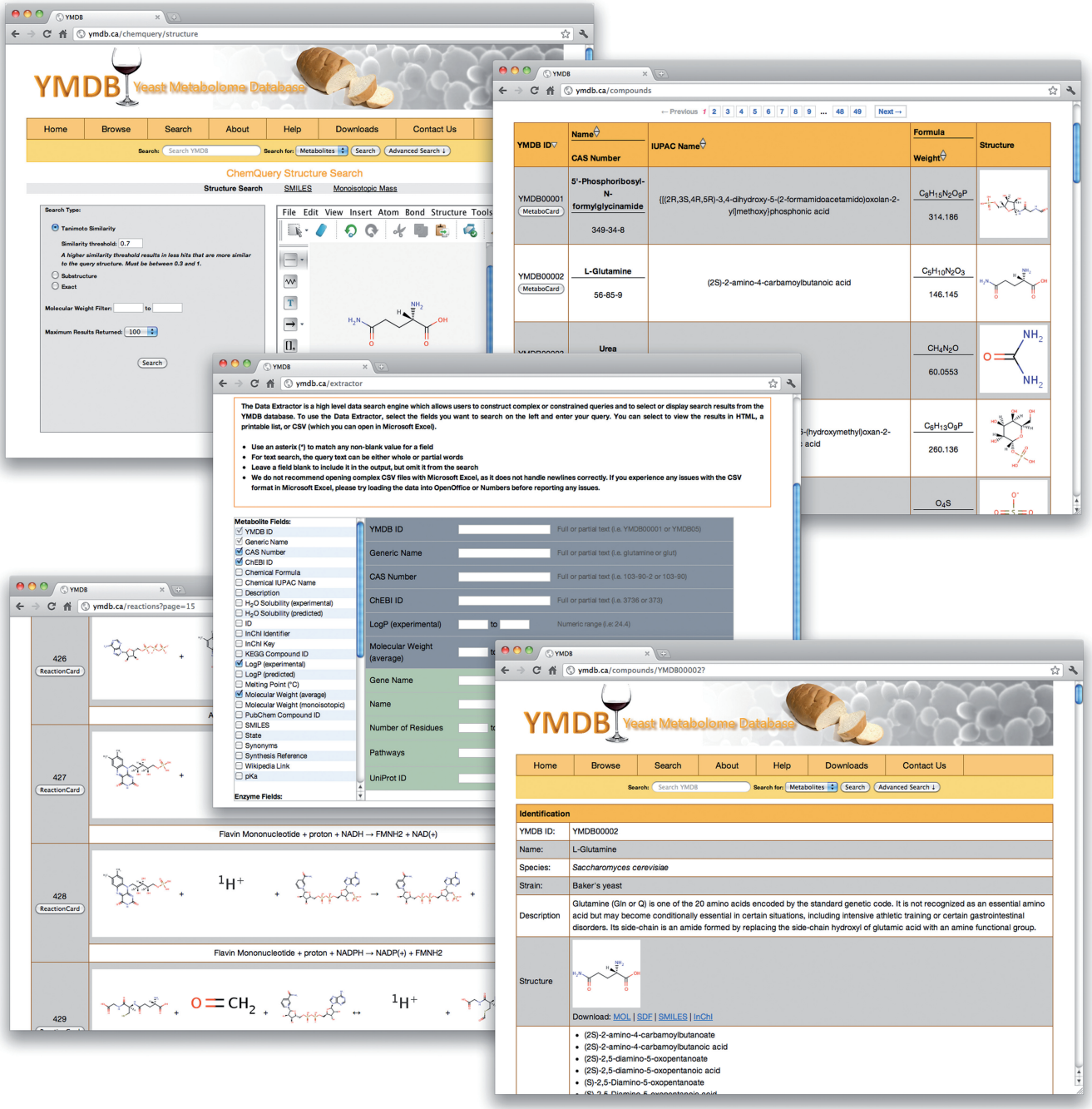

Figure 1. A screenshot montage of the YMDB showing several of the YMDB's search and data display tools describing the metabolite L-Glutamine. Not all fields are shown.

spectral search utilities allow both pure compounds and mixtures of compounds to be identified from their MS or NMR spectra via peak matching algorithms that were developed in-house (28,29).

Adjacent to the 'Search' menu, the 'About' pull-down menu contains information on the YMDB database, recent news or updates, links to other databases, data sources and database statistics. The 'Help' pull-down menu provides general documentation on database definitions, data field types and data field sources. It also contains information on experimental methods (for metabolite concentration measurements performed by our lab for the YMDB), details on how to cite YMDB, as well as a tutorial on how to use YMDB's advanced text search utilities. Finally the 'Download' menu contains downloadable data for all YMDB chemical structures (as Structure Data Format (SDF) files), all enzyme/protein sequences (in FASTA format), as well as complete flat file data sets of the current YMDB release in JSON format.

DATABASE IMPLEMENTATION

YMDB employs a Ruby on Rails (version: 3.09)-based front-end attached to a sophisticated MySQL relational database (version: 5.0.77) at its back-end. All data are entered directly through a custom-built web interface with each YMDB MetaboCard having an edit page, which allows database curators to manually make changes to YMDB entries. The public user interface and the internal database both read from the same database.

All structures in the YMDB are stored in a centralized structure hub. This hub is a RESTful web resource that automatically stores and updates chemical properties such as molecular weight, solubility and logP. Additionally, the hub renders the structure images and thumbnails visible on the public YMDB site. The centralized nature of this structure hub helps to maintain consistency for all structures stored in YMDB. Whenever a structure is changed or updated, all properties are automatically recalculated and made available on the public site at http://www.ymdb.ca.

QUALITY ASSURANCE, COMPLETENESS AND CURATION

The same quality assurance, quality control and data compilation procedures implemented during the development of HMDB, T3DB and DrugBank were used in the development of YMDB. In particular, the compounds in YMDB were identified using a combination of methods, including manual literature surveys, text mining of on-line journals or abstracts and data mining of other electronic databases. Literature sources included specialty journals on metabolomics, food composition and analysis, systems biology, analytical chemistry and textbooks on wine and beer chemistry. All primary metabolites had to have at least two databases confirm their existence and inclusion (with evidence that the necessary enzymes or pathways are present), whereas all secondary metabolites (such as those found in wine or beer) were required to have a traceable literature/experimental reference. For many secondary yeast metabolites the relevant starting compounds, reactions, pathways and catalyzing enzymes are not yet known. Hopefully, with time and improved technology, this information will become available. With many yeast secondary metabolites it is sometimes difficult to know if the compound was present in the media (wort or grape must) prior to fermentation or whether it arose as a consequence of fermentation. For those compounds where there was some ambiguity regarding their source (plant versus yeast), we attempted to cross-check our findings through multiple literature sources in order to exclude possible grape, hops or barley metabolites.

For those yeast metabolites found to match to previously existing entries from either the HMDB or CCDB, only the chemical data fields were imported into the YMDB (except the compound description which was manually edited to include or remove organism-specific references). The biological data for these HMDB/CCDB imported compounds was generated *de novo* since yeast biology is very different than *E. coli* or human biology. In order to ensure both completeness and correctness, each metabolite record entered into the YMDB was reviewed and validated by a member of the curation team after being annotated by another member. Other members of the curation group routinely performed additional spot checks on each entry. Several software packages including text-mining tools, chemical parameter calculators and protein annotation tools were developed, modified and used to aid in data entry and data validation. One particular program, BioSpider (30), was used extensively to acquire routine, machine retrievable or easily calculated/verifiable chemical data on metabolites. To facilitate and monitor the data entry process, all of YMDB's data is entered into a centralized, password-controlled database, allowing all changes and edits to the YMDB to be monitored, time-stamped and automatically transferred.

CONCLUSIONS

To summarize, the YMDB is a richly annotated, web-accessible 'metabolomics' database that brings together quantitative chemical, physical and biological data about nearly 2000 *S. cerevisiae* metabolites. Relative to other yeast metabolism/pathway databases, YMDB has between 2–3× more metabolites and 5–10× more data. The YMDB also uniquely contains detailed information on hundreds of secondary metabolites that are critically important to the food, beverage, chemical and biofuel industry. Among the other distinguishing features of YMDB are: (i) the breadth and depth of its annotations (>80 data fields); (ii) the large number of hyperlinks and references to other resources; (iii) the availability of detailed compound descriptions; (iv) the inclusion of thousands of reference NMR and MS spectral data; (v) the inclusion of intra- and extracellular metabolite concentration data; (vi) the quantity of biological and biochemical information included in each compound entry and (vii) the support for queries by text, chemical structure, spectra, molecular weight and gene/protein sequence. Owing to these unique characteristics, we believe the YMDB fills an important niche in yeast biology as it addresses not only the specialized analytical needs of metabolomics researchers, but also the interests of molecular biologists, systems biologists, the industrial fermentation industry, as well as the beer, wine and spirit industries.

While the YMDB certainly fills an important niche for yeast metabolomics, it is also a work in progress. As with many areas in metabolomics, new compounds are constantly being discovered, new concentrations are being reported, new pathways/reactions are being elucidated and new metabolite functions are being determined. So long as our resources permit, we intend to continue to update and enhance the YMDB as this new information is published or acquired.

FUNDING

The Canadian Institutes of Health Research (CIHR); Agriculture and Agri-Food Canada (Agriculture

Bioproducts Innovation Program); Genome Alberta, a division of Genome Canada. Funding for open access charge: Genome Canada.

Conflict of interest statement. None declared.

REFERENCES

1. Weckwerth,W. (2010) Metabolomics: an integral technique in systems biology. *Bioanalysis*, **2**, 829–836.
2. Wishart,D.S. (2007) Current progress in computational metabolomics. *Brief. Bioinform.*, **8**, 279–293.
3. Wohlgemuth,G., Haldiya,P.K., Willighagen,E., Kind,T. and Fiehn,O. (2010) The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, **26**, 2647–2648.
4. Xia,J., Psychogios,N., Young,N. and Wishart,D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.*, **37**, W652–W660.
5. Wishart,D.S., Knox,C., Guo,A.C., Eisner,R., Young,N., Gautam,B., Hau,D.D., Psychogios,N., Dong,E., Bouatra,S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.
6. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
7. Sundararaj,S., Guo,A., Habibi-Nazhad,B., Rouani,M., Stothard,P., Ellison,M. and Wishart,D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. *Nucleic Acids Res.*, **32**, D293–D295.
8. Lim,E., Pon,A., Djoumbou,Y., Knox,C., Shrivastava,S., Guo,A.C., Neveu,V. and Wishart,D.S. (2010) T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res.*, **38**, D781–D786.
9. Frolkis,A., Knox,C., Lim,E., Jewison,T., Law,V., Hau,D.D., Liu,P., Gautam,B., Ly,S. *et al.* (2010) SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res.*, **38**, D480–D487.
10. Scalbert,A., Brennan,L., Fiehn,O., Hankemeier,T., Kristal,B.S., van Ommen,B., Pujos-Guillot,E., Verheij,E., Wishart,D. and Wopereis,S. (2009) Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, **5**, 435–458.
11. Jernigan,D.H. (2004) The global alcohol industry: an overview. *Addiction*, **104(Suppl. 1)**, 6–12.
12. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* Life with 6000 genes. *Science*, **274**, 563–567.
13. Costanzo,M. and Boone,C. (2009) SGAM: an array-based approach for high-resolution genetic mapping in *Saccharomyces cerevisiae*. *Methods Mol. Biol.*, **548**, 37–53.
14. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L., Toufighi,K., Mostafavi,S. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
15. Engel,S.R., Balakrishnan,R., Binkley,G., Christie,K.R., Costanzo,M.C., Dwight,S.S., Fisk,D.G., Hirschman,J.E., Hitz,B.C., Hong,E.L. *et al.* (2010) Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res.*, **38**, D433–D436.
16. Hodges,P.E., McKee,A.H., Davis,B.P., Payne,W.E. and Garrels,J.I. (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, **27**, 69–73.
17. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martínez,J., Pérez-Ortín,J.E. *et al.* (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.
18. Robinson,M.D., Grigull,J., Mohammad,N. and Hughes,T.R. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.
19. Herrgård,M.J., Swainston,N., Dobson,P., Dunn,W.B., Arga,K.Y., Arvas,M., Blüthgen,N., Borger,S., Costenoble,R., Heinemann,M. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, **26**, 1155–1160.
20. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
21. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
22. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
23. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
24. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcántara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
25. Weininger,D. (1988) SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–38.
26. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
27. Horai,H., Arita,M., Kanaya,S., Nihei,Y., Ikeda,T., Suwa,K., Ojima,Y., Tanaka,K., Tanaka,S., Aoshima,K. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.
28. Xia,J., Bjorndahl,T.C., Tang,P. and Wishart,D.S. (2008) MetaboMiner – Semi-automated Identification of Metabolites from 2D NMR Spectra of Complex Biofluids. *BMC Bioinformatics*, **9**, 507.
29. Dworzanski,J.P., Snyder,A.P., Chen,R., Zhang,H., Wishart,D.S. and Li,L. (2004) Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. *Anal. Chem.*, **76**, 2355–2366.
30. Knox,C., Shrivastava,S., Stothard,P., Eisner,R. and Wishart,D.S. (2007) BioSpider: a web server for automating metabolome annotations. *Pac. Symp. Biocomput.*, 145–156.